International Journal of Research in Computer Science
eISSN 2249-8265 Volume 2 Issue 2 (2012) pp. 29-32

www.ijorcs.org

DUPLICATE DETECTION OF RECORDS IN QUERIES USING CLUSTERING

M.Anitha[1], A.Srinivas[2], T.P.Shekhar[3], D.Sagar[4]

** Sree Chaitanya College Of Engineering (SCCE), Karimnagar, India*

Abstract: *The problem of detecting and eliminating duplicated data is one of the major problems in the broad area of data cleaning and data quality in data warehouse. Many times, the same logical real world entity may have multiple representations in the data warehouse. Duplicate elimination is hard because it is caused by several types of errors like typographical errors, and different representations of the same logical value. Also, it is important to detect and clean equivalence errors because an equivalence error may result in several duplicate tuples. Recent research efforts have focused on the issue of duplicate elimination in data warehouses. This entails trying to match inexact duplicate records, which are records that refer to the same real-world entity while not being syntactically equivalent. This paper mainly focuses on efficient detection and elimination of duplicate data. The main objective of this research work is to detect exact and inexact duplicates by using duplicate detection and elimination rules. This approach is used to improve the efficiency of the data.*

Keywords: *Data Cleaning, Duplicate Data, Data Warehouse, Data Mining*

I. INTRODUCTION

Data warehouse contains large amounts of data for data mining to analyze the data for decision making process. Data miners do not simply analyze data, they have to bring the data in a format and state that allows for this analysis. It has been estimated that the actual mining of data only makes up 10% of the time required for the complete knowledge discovery process [3]. According to Jiawei, the precedent time consuming step of preprocessing is of essential important for data mining. It is more than a tedious necessity: The techniques used in the preprocessing step can deeply influence the results of the following step, the actual application of a data mining algorithm [6]. Hans-peter stated as the role of the impact on the link of data preprocessing to data mining will gain steadily more interest over the coming years. Preprocessing is one of the fourth future trend and major issues in data mining over the next years [7].

In data warehouse, data is integrated or collected from multiple sources. While integrating data from multiple sources, the amount of the data increases and as well as data is duplicated. Data warehouse may have terabyte of data for the mining process. The preprocessing of data is the initial and often crucial step of the data mining process. To increase the accuracy of the mining result one has to perform data preprocessing because 80% of mining efforts often spend their time on data quality. So, data cleaning is very much important in data warehouse before the mining process. The result of the data mining process will not be accurate because of the data duplication and poor quality of data. There are many existing methods available for duplicate data detection and elimination. But the speed of the data cleaning process is very slow and the time taken for the cleaning process is high with large amount of data. So, there is a need to reduce time and increase speed of the data cleaning process as well as need to improve the quality of the data.

There are two issues to be considered for duplicate detection: *Accuracy and Speed.* The measure of accuracy in duplicate detection depends on the number of false negatives (duplicates you did not classify as such) and false positives (non-duplicates which were classified as duplicates) [12].

In this research work, a duplicate detection and elimination rule is developed to handle any duplicate data in a data warehouse. Duplicate elimination is very important to identify which duplicate to retain and duplicate is to be removed. The main objective of this research work is to reduce the number of false positives, to speed up the data cleaning process reduce the complexity and to improve the quality of data. A high quality, scalable duplicate elimination algorithm is used and evaluated it on real datasets from an operational data warehouse to achieve objective.

II. RECORD MATCHING OVER QUERY RESULTS

A. First Problem Definition

Our focus is to find the matching status among the records and to retain the non duplicate records. Then, the goal is to cluster the matched records using fuzzy ontological document clustering.

B. Element Identification

Supervised learning methods use only some of the fields in a record for identification. This is the reason for query results obtained using supervised learning to contain duplicate records. Unsupervised Duplicate Elimination (UDE) does not suffer from these types of user reference problems. A preprocessing step called exact matching is used for matching relevant records. It requires the data format of the records to be the same. So, the exact matching method is applicable only for the records from the same data source. Element identification thus merges the records that are exactly the same in relevant matching fields.

C. Ontology matching

The term *Ontology* is derived from the Greek words '*onto'* which means *being* and '*logia'* which means *written* or *spoken disclosure*. In short, it refers to a specification of a conceptualization.

Ontology basically refers to the set of concepts such as things, events and relations that are specified in some way in order to create an agreed-upon vocabulary for exchanging information. Ontologies can be represented in textual or graphical formats. Usually, graphical formats are preferred for easy understandability. Ontologies with a large knowledge base [5] can be represented in different forms such as hierarchical trees, expandable hierarchical trees, hyperbolic trees, etc. In the expandable hierarchical tree format, the user has the freedom to expand only the node of interest and leave the rest in a collapsed state [2]. If necessary, the entire tree can be expanded to get the complete knowledge base. This type of format can be used only when there are a large number of hierarchical relationships. Ontology matching is used for finding the matching status of the record pairs by matching the record attributes.

III. SYSTEM METHODOLOGY

A. Unsupervised Duplicate Elimination

UDE employs a similarity function to find field similarity. We use similarity vector to represent a pair of records.

Input: Potential duplicate vector set P Non-duplicate vector set N

Output: Duplicate vector set D

C_1 : A classification algorithm with adjustable parameters W that identifies duplicate vector pairs from P C_2 : a supervised classifier, SVM

Algorithm:

1. D = φ
2. Set the parameters W of C_1 according to N
3. Use C_1 to get a set of duplicate vector pairs d_1 and f from P and N
4. P = P- d_1
5. while | d_1 |≠ 0
6. N' = N - f
7. D = D + d_1 + f
8. Train C_2 using D and N'
9. Classify P using C_2 and get a set of newly identified duplicate vector pairs d_2
10. P = P - d_2
11. D =D + d_2
12. Adjust the parameters W of C_1 according to N' and D
13. Use C_1 to get a new set of duplicate vector pairs d_1 and f from P and N
14. N=N'
15. Return D

Figure 1: UDE Algorithm

B. Certainty factor

In the existing method of duplicate data elimination [10], certainty factor (CF) is calculated by classifying attributes with distinct and missing value, type and size of the attribute. These attributes are identified manually based on the type of the data and the most important of data in that data warehouse. For example, if name, telephone and fax field are used for matching then high value is assigned for certainty factor. In this research work, best attributes are identified in the early stages of data cleaning. The attributes are selected based on the specific criteria and quality of the data. Attribute threshold value is calculated based on the measurement type and size of the data. These selected attributes are well suited for the data cleaning process. Certainty factor is assigned based on the attribute types. This is shown in the following table.

Table 1: Classification of attribute types

S. No	Key Attribute	Distinct values	Missing values	Size of data	Types of data
1	√			√	√
2	√			√	
3	√				√
4		√	√	√	√
5		√	√	√	
6		√	√		√
7		√		√	√
8		√	√		
9		√			√
10		√		√	

Rule 1: certainty factor 0.95 (No. 1 and No. 4)

- Matching key field with high type and high size
- And matching field with high distinct value, low missing value, high value data type and matching field with high range value

Rule 2: certainty factor 0.9 (No. 2 and No. 4)

- Matching key field with high range value
- And matching field with high distinct value, low missing value, and matching field with high range value

Rule 3: certainty factor 0.9 (No. 3 and No. 4)

- Matching key field with high type
- And Matching field with high distinct value, low missing value, high value data type and matching field with high range value

Rule 4: certainty factor 0.85 (No. 1 and No. 5)

- Matching key field with high type and high size
- And matching field with high distinct value, low missing value and high range value

Rule 5: certainty factor 0.85 (No. 1 and No. 5)

- Matching key field and high size
- And matching field with high distinct value, low missing value and high range value

Rule 6: certainty factor 0.85 (No. 2 and No. 5)

- Matching key field with high type
- And matching field with high distinct value, low missing value and high range value

Rule 7: certainty factor 0.85 (No. 1 and No. 6)

- Matching key field with high size and high type
- And matching field with high distinct value, low missing value and high value data type

Rule 8: certainty factor 0.8 (No. 3 and No. 7)

- Matching key field with high type
- And Matching field with high distinct value, high value data type and high range value

Rule 09: certainty factor 0.75 (No. 2 and No. 8)

- Matching key field with high size
- And matching field with high distinct value and low missing value

Rule 10: certainty factor 0.75 (No. 3 and No. 8)

- Matching key field with high type
- And matching field with high distinct value and low missing value

Rule 11: certainty factor 0.7 (No. 1 and No. 9)

- Matching key field with high type and high size
- And matching field with high distinct value and high value data type

Rule 12: certainty factor 0.7 (No. 2 and No. 9)

- Matching key field with high size
- And matching field with high distinct value and high value data type

Rule 13: certainty factor 0.7 (No. 3 and No. 9)

- Matching key field with high type
- And matching field with high distinct value and high value data type

Rule 14: certainty factor 0.7 (No. 1 and No. 10)

- Matching key field with high type and high size
- And matching field with high distinct value and high range value

Rule 15: certainty factor 0.7 (No. 2 and No. 10)

- Matching key field with high size
- And matching field with high distinct value and high range value

Rule 16: certainty factor 0.7 (No. 3 and No. 10)

- Matching key field with high type
- And matching field with high distinct value and high range value

S No.	Rules	Certainty Factor (CF)	Threshold value (TH)
1	{TS}, {D, M, DT, DS}	0.95	0.75
2	{T, S}, {D, M, DT, DS}	0.9	0.80
3	{TS, T, S}, {D, M, DT}, {D, M, DS}	0.85	0.85
4	{TS, T, S}, {D, DT, DS}	0.8	0.9
5	{TS, T, S}, {D, M}	0.75	0.95
6	{TS, T, S}, {D, DT}, {D, DS}	0.7	0.95

TS – Type and Size of key attribute
T – Type of key attribute
S – Size of key attributes
D – Distinct value of attributes
M – Missing value of attributes
DT – Data type of attributes
DS – Data size of attributes

Duplicate records are identified in each cluster to identify exact and inexact duplicate records. The duplicate records can be categorized as match, may be match and no-match. Match and may be match duplicate records are used in the duplicate data elimination rule. Duplicate data elimination rule will identify the quality of the each duplicate record to eliminate poor quality duplicate records.

- Calculate the similarity of the documents matched in main concepts (*X*mc) and the similarity of the documents matched in detailed descriptions (*X*dd).
- Evaluate *X*mc and *X*dd using the rules to derive the corresponding memberships.
- Compare the memberships and select the minimum membership from these two sets to represent the membership of the corresponding concept (high similarity, medium similarity, and low similarity) for each rule.
- Collect memberships which represent the same concept in one set.
- Derive the maximum membership for each set, and compute the final inference result.

C. Evaluation Metric

The overall performance can be found using precision and recall where

$$Precision = \frac{\text{Number of correctly identified duplicate pairs}}{\text{Number of all identified duplicate pairs}}$$

$$Recall = \frac{\text{Number of correctly identified duplicate pairs}}{\text{Number of true duplicate pairs}}$$

The classification quality is evaluated using F-measure which is the harmonic mean of precision and recall

$$F - measure = \frac{2(precision)(recall)}{precision+recall}$$

IV. CONCLUSION

Deduplication and data linkage are important tasks in the pre-processing step for many data mining projects. It is important to improve data quality before data is loaded into data warehouse. Locating approximate duplicates in large data warehouse is an important part of data management and plays a critical role in the data cleaning process. In this research wok, a framework is designed to clean duplicate data for improving data quality and also to support any subject oriented data.y

In this research work, efficient duplicate detection and duplicate elimination approach is developed to obtain good result of duplicate detection and elimination by reducing false positives. Performance of this research work shows that the time saved significantly and improved duplicate results than existing approach.

The framework is mainly developed to increase the speed of the duplicate data detection and elimination process and to increase the quality of the data by identifying true duplicates and strict enough to keep out false-positives. The accuracy and efficiency of duplicate elimination strategies are improved by introducing the concept of a certainty factor for a rule. Data cleansing is a complex and challenging problem. This rule-based strategy helps to manage the complexity, but does not remove that complexity. This approach can be applied to any subject oriented data warehouse in any domain.

V. REFERENCES

[1] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, “Robust and Efficient Fuzzy Match for Online Data Cleaning,” Proc. ACM SIGMOD, pp. 313-324, 2003.

[2] Kuhanandha Mahalingam and Michael N.Huhns, “Representing and using Ontologies”,USC-CIT Technical Report 98-01.

[3] Weifeng Su, Jiying Wang, and Federick H.Lochovsky, ” Record Matching over Query Results from Multiple Web Databases” IEEE transactions on Knowledge and Data Engineering, vol. 22, N0.4,2010.

[4] R. Ananthakrishna, S. Chaudhuri, and V. Ganti, “Eliminating Fuzzy Duplicates in Data Warehouses. VLDB”, pages 586-597, 2002.

[5] Tetlow.P,Pan.J,Oberle.D,Wallace.E,Uschold.M,Kendall .E,”Ontology Driven Architectures and Potential Uses of the Semantic Web in Software Engineering”,W3C,Semantic Web Best Practices and Deployment Working Group,Draft(2006).

[6] Ji-Rong Wen, Fred Lochovsky, Wei-Ying Ma, **“**Instance-based Schema Matching for Web Databases by Domain-specific Query Probing”, Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004.

[7] Amy J.C.Trappey, Charles V.Trappey, Fu-Chiang Hsu,and David W.Hsiao, “A Fuzzy Ontological Knowledge Document Clustering Methodology”,IEEE Transactions on Systems,Man,and Cybernetics-Part B:Cybernetics,Vol.39,No.3,june 2009.